Metadata and the Semantic Web

United States Army
Information Systems Engineering Command

By: Mr. Trace Gunsch, Critical Skills Expert for Emerging Technology mailto:gunscht@hqisec.army.mil and Mr. Stuart Engle, Multimedia Engineer.

An unmanned vehicle crests the top of a hill and stops. Sensors survey the terrain, collecting sonar, imagery, and thermal data. Each instrument sends data with generated metadata, sorted into 16 area-specific objects by a software agent. The agent then triggers another software agent to assimilate the objects into a series of images and correlate them with a map. Modeling programs in the vehicle's central processor process the objects for recognition. A situation profile, including a text report, is transmitted to an operator twenty miles away. The report: `Thirteen enemy tanks (T-70's) and three anti-aircraft track-mounted vehicles (ZSU-23-4)are moving southwest (12 km/hr).` *As the report is sent, motion sensors detect one tank halting, signifying possible detection. An alert passes to the behavior agent, which prompts the vehicle into reverse and sends it back down the hill beyond the range of fire.*

INTRODUCTION

The software technology utilized in this advanced, unmanned scout can be applied to the existing Internet, transforming it from a display monitor to a network of information processors, called the **Semantic Web**. The technologies of the Semantic Web enable machines to make more sense of the Web, with the result of making the Web more useful for humans.

The abundance of information available on the Internet today is much like the aggregation of information amassed by multiple sources of the military when threatened. In both cases, the challenge to identify, assimilate, and process the right information is directly related to success. Computers provided the information, but were not empowered to transfer it globally, or process it to make decisions. When new and legacy systems from all service branches can properly exchange information, perform functions, and serve applications with **minimal human manipulation,** decision superiority will be assured.

> "The Joint force must be able to take advantage of superior information converted to superior knowledge to achieve "decision superiority" – better decisions arrived at and implemented faster than an opponent can react, or in a non-combat situation, at a tempo that allows the force to shape the situation or react to changes and accomplish its mission."
>
> **-- Joint Vision 2020 General Henry H. Shelton**

The United States Army has taken a first step toward semantic transformation. The Army Knowledge Online (AKO) portal is the vehicle for becoming "…a network-centric, knowledge-based force" (Army Knowledge Management Strategic Plan, 7). The immediate effect has been to link all Army personnel with an account on AKO, thus creating a network for communication and information dissemination. To enable semantic transformation, AKO leaders wisely insisted on enforcing the use of open, standards-based Information Technology (IT) architectures that included an initiative for "…an evolving strategy and a set of tasks for the effective and well-coordinated use of Extensible Markup Language (XML) to support Army functions" (Army Knowledge Management Strategic Plan, 29) These functions, from the sustaining base to the tactical arena, will be performed on-line through a Semantic Web, bringing decision superiority to every facet of the Army's operations.

BACKGROUND

The Internet originated as a means of sending and retrieving text among research scientists, but with the emergence of Hypertext Mark-Up Language (HTML) quickly grew into a medium for sharing information in an increasing variety of formats, including images, audio, and video. While supplying nearly universal access to a global warehouse of information, it does little to enable users to find **context-specific** information. Websites and links are not posted to the Internet with a prescribed organizational scheme, so the capacity to retrieve the right information at the right time is largely a matter of luck, not design.

If today's Internet, or the underlying Transmission Control Protocol/Internet Protocol (TCP/IP), had been designed with a robust application of Standard Generalized Markup Language (SGML) that included metadata and extensibility, networked machines would already exchange and process information, rather than simply display it. Specifically, TCP/IP, metadata, and markup languages enable computers to perform the following tasks:

a. Share data across platforms using a network.

b. Apply rules to the arrangement of the data's structure to process logical operations.

With the existing Internet, a **user** sifts through the results of a search looking for specifics. As defined earlier, the Semantic Web will enable **machines** to sift for context specific information, as prescribed by the user, and present a focused set of results. The machines do more between themselves without human interaction, thus making humans more efficient. Applying the utility of a mark up language begins with metadata and the hierarchical structuring of tags.

METADATA

Information about information is generally known as metadata. On the web, metadata is machine-understandable information about web resources. It can be contained within a web page's source code, transferred along with the web page as header information, or referenced in a separate document. The relationship of metadata to its information allows a machine to identify, regard, and process it according to a programmed intent.

For a simple analogy, an instance of metadata could be a list of items on an aisle in a warehouse. The list identifies the aisle's contents and can be further categorized to enable efficient search and retrieval of a specific item. The list also helps a patron decide whether the aisle contains anything worth pursuing without having to physically walk through it.

No such list exists for the Internet. The closest provision is the <META> tag, which labels a page for inspection by a search engine. Web authors type in words that indicate the page's content, but search engines compare every meta-tag on every page to the text of the query and return all matches. No shortcuts or filters like author, subject, or year exist to reduce the number of instances for comparison, or define the context for a specifically required match.

The XML provides a mechanism for browsers and web applications to define, recognize, and manipulate information using a custom set of tags. Creating and interpreting tags, such as the <img> tag for pictures, allows information to be transmitted and displayed across the web. Metadata can be tagged and nested to create a tree-like structure that conveys position in a hierarchy, thereby enabling the computer to reference the content of the metadata by position. Hence, metadata can be assigned to information on the web, such as a database, and be interpreted as a multi-dimensional index to the fields and records within the database, greatly enhancing the utility of its content. It also provides sufficient relational information for the use of ruses in decision-making operations.

DEFINING TAGS

With the emergence of XML, web authors began writing their own unique tags in HTML documents. To define the task, a reference to a separate file, called a Document Type Definition (DTD) was included. When the browser loads the page, the DTD is retrieved so that the tags can be interpreted. While this works well for individual web authors, communities of professionals prefer to share, or standardize, tags useful for defining particular properties of data used in their work. For example, the publishing industry created a set of metadata tags called the Dublin Core. The group associated different types of information about publications to the tags referenced in their DTD. These tags included categories such as <CREATOR>, <SUBJECT>, <PUBLISHER>, <DATE>, <FORMAT>, <LANGUAGE>, and others. In effect, the tags serve to categorize information. Searches can be more direct and return more concise results.

While referencing an arbitrary set of tags is convenient, it generates a different problem between groups. The publishing industry defines creator as an author, while the motion picture industry considers a creator to be a director. The resulting ambiguity requires clarification through an explicit definition of the directory; for example, by occupation. Two solutions emerged for explicit definitions: **namespaces** and **Resource Description Frameworks** (RDF).

XML standards require that entities be referenced properly before parsing. Namespaces serve to reference the entity, whether it is a piece of information, a document, or link by specifying it as a Universal Resource Identifier (URI). These references provide further information about custom and/or standard tags used in the file. Conventions for syntax and inheritance apply, and resolve any ambiguity caused by variant uses of a common set, or structure, of tags.

RDFs also use URIs but as sets of three, so that a relationship between two entities can be defined. For example, the sentence, **John is the brother of Jim** contains three parts, with each part referenced by a URI. A machine would read it as URI A (John) – URI B (the brother of) – URI C (Jim). The RDF then goes beyond XML syntax by establishing classes (the family John and him belong to), properties (resources and characteristics of the family members), and the relationships among them (parent-child). This arrangement of entities, or **schema**, defines a scalable, hierarchical structure that can be read and processed at the machine level, including inferences about data processed on the pages. RDFs and schemas are sufficient for developing metadata applications. The World Wide Web Consortium has organized a working group to develop a Schema Definition Language to define metadata systems and enable them to inherit metadata definitions from other schemas to extend its capabilities.

ONTOLOGIES

In the warehouse analogy, an RDF would serve to specify the height and width of shelves and equip the items with bar coding so that robotic forklifts (agents) could maneuver, organize, locate, and retrieve items. This works well as long as the robotic forklift stays in the area described by that particular RDF or receives appropriate instructions from a referenced schema. In reality, distinctly different communities have created unique RDFs that work with file types other than text. This requires another level of organization and syntax to coordinate semantic interoperability between different communities and the agents operating within them.

The requirement for such a level, called ontology, is most conspicuous when applied to multimedia file formats. Searching a video file for text is useless without metadata attached. Other aspects of video, such as duration, compression scheme, and segmentation require unique metadata tags. Automating the generation of metadata during file production enables content access and manipulation most conveniently. The Motion Picture Experts Group (MPEG) first created an ontology called the Multimedia Content Description Interface, defined as MPEG-7. These standards ". . .enable machines to generate and understand audiovisual descriptions for retrieval, categorisation, and filtering purposes. Significant progress has been made on automatic segmentation, scene-change detection, and the recognition and detection of low-level features for multimedia content." (Hunter, section 1)

Making the metadata from the MPEG-7 standard interoperable with the educational community and/or museum professionals required another ontology. The MPEG met again to develop an ontology, MPEG-21. In "Enhancing the Semantic Interoperability of Multimedia through a Core Ontology," Jane Hunter explains:

"The objectives of MPEG's latest initiative, MPEG-21 [2] (ISO/IEC 18034-1), are to:

- Provide a vision for a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and devices to meet the needs of all users;

- Facilitate the integration of components and standards in order to harmonise technologies for the creation, management, manipulation, transport, distribution and consumption of content;
- Provide a strategy for achieving a multimedia framework by the development of specifications and standards based on well-defined functional requirement through collaboration with other bodies."

Development for these goals led to anticipation of a **core ontology** capable of incorporating various RDFs and other XML schemas, including MPEG-7 and MPEG-21. While this anticipation has yet to yield formal standards, it represents the foundational level for building and integrating existing sets of metadata tags, whether defined by DTDs, XML schemas, RDFs or ontologies, into an interoperable, network-centric knowledge base. This leverages web functionality for more search, retrieval, and processing between machines and media without human involvement. It also suggests using metadata and markup language technologies as a means for communicating and controlling devices. This is the enabling structure for the transformation of the Internet to a Semantic Web.

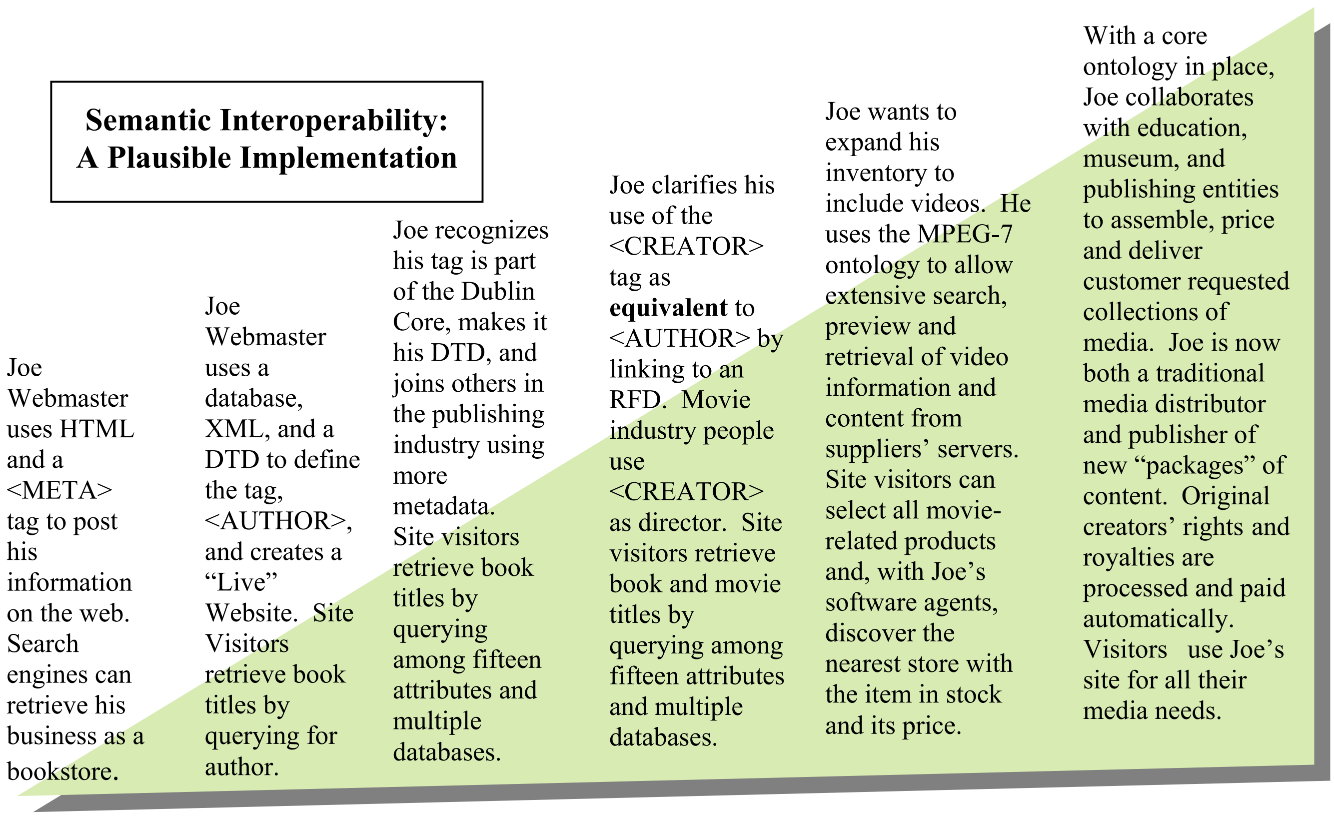

Figure 1. Semantic Web Technology Progression

GOVERNMENT IMPLEMENTATIONS OF XML

Several federal agencies have begun implementing XML schemas, including Defense Advanced Research Project Agency (DARPA) with their DARPA Agent Markup Language (DAML) and National Institute of Standards and Technology's (NIST) Intelligent Systems Division. The following paragraphs describe some of these implementations.

The DARPA DAML Concept

Before AKO mandated the use of SML in 2000, military research with metadata became evident in 1998 at DARPA. Beginning with the concept of Control of Agent-Based Systems (CoABS), DARPA sought ". . .a fundamentally new kind of software technology for automating our processes." [Hendler, DARPATech '99 speech transcript] This led to the development of a new application of SGML, termed DARPA Agent Markup Language (DAML). In the presentation "Agent Based Computing, (slide 13) DARPA's Information Systems Office explains that the goal of this language, implemented at the ontology level, was to "…enable software agents to identify, communicate with, and understand other software agents dynamically (i.e., on the fly at run time, not built in at development time)." These software agents would act as interpreters between sensors and legacy and evolving systems, interconnected by an interoperability grid, and as programs for analysis. Exploitation of metadata required DAML, for three reasons:

a. Metadata content is assigned to systems submitting information. This signifies what system sent the information and, due to its relative position in the hierarchy, how the agent can use the information.

b. Defining tags in an object-oriented hierarchy renders data as elements subject to class definitions, object properties, inheritance and other characteristics that enable advanced processing. Software tools can subsequently be created to facilitate the development of agents, capable of specific functions.

c. Data from sensors, systems, and models equipped with DAML structure can be transmitted across a "software grid" on a ubiquitous TCP/IP (Internet) hardware architecture.

DARPA plans to complete DAML v. 2 in mid 2003 and follow up with the second version of the Taskable Agent Software Kit (TASK) in mid 2004. These releases should propel the achievement of semantic interoperability into tangible benefits.

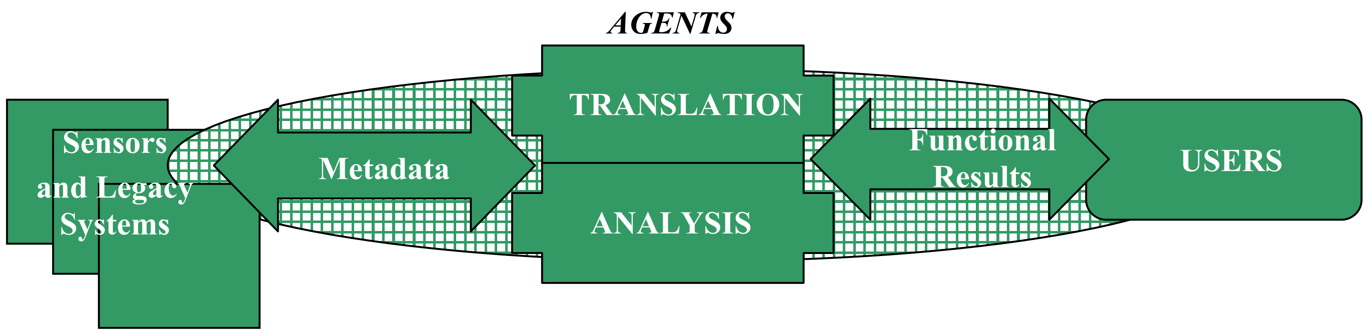

Figure 2: DAML and Semantic Web Implementation

NIST Conceptual Implementation

Several agencies are implementing metadata systems. The National Institute of Standards and Technology's Intelligent Systems Division created the vehicle described at the beginning of the paper. It is equipped with sensors, intelligent agent software, and an interface to communicate with a remote user. **Sensory processing** collects data from sensors and does analysis to filter, recognize, compute, group and window data. The data is passed, through semantic interoperability techniques, to the **world modeling area** where it is processed into images, maps, states, attributes, values, entities and events. The **behavior generation domain**, comprised of intelligent agent software, receives the constructs and applies semantic reasoning to generate planners and actuators before triggering responses in the physical environment.

Metadata in the Army

Assigning metadata and XML schemas to make all real world systems interoperable and accessible as web-based applications remains a huge challenge. Given the digital nature of the information and "anytime, anywhere" utility of a network, the inclusion of metadata as described above enables a host of applications across many domains. Practical applications of the use of metadata, interoperability, and agents present new and powerful benefits to the Army. The following situations illustrate a few current implementations and suggest future uses as well.

a. Military Intelligence: the All Source Analysis System – Light is a laptop intelligence processor that uses XML as a method of porting and processing data between systems on a local area network (LAN), allowing the transfer of overlays from different mapping utilities and 3D renderings of data.

b. Digital Resources: the U.S. Army Publishing Agency has begun logging tags for use with digital files. Metadata tags mark source information into modules, categorized by function, and grouped as work packages. Search, retrieval, and reuse are greatly enhanced.

c. Distance Learning: The Training and Doctrine Command (TRADOC) requires Army Distance Learning materials to have an .xml file with Information Management System (IMS) tags appended before accepting them. This enables training materials to be dynamically assembled for the needs of a particular soldier or unit. If the IMS metadata vocabulary were incorporated into an ontology that included MPEG-21, this functionality could be extended to include resources of all types, including printed material, training modules, news reports, and multimedia.

Potential Military Uses of Metadata

a. Automated "loading" of imagery: Map applications on intelligence processors could download satellite images from the National Imagery and Mapping Agency (NIMA) web site. Metadata from the image files could be imported to automate the alignment of the image with other geographically-referenced, map data.

b. Operational Orders (OPORD) for an echelon could be submitted to intelligent agents for analysis and comparison to current data on subordinate units. The OPORD could be revised, based on a "best-fit" match of units to areas and anticipated threats.

c. Requests for information from systems at lower echelons could contain metadata that identifies itself to the system at the higher echelon. The metadata would include formatting information to facilitate the incorporation of the response data.

d. A central repository could maintain templates of enemy orders of battle. Brigade intelligence analysts could then submit his/her current situational data to intelligent agents for comparison. The central repository could return plausible enemy courses of action for consideration.

Warning: Potential Metadata "Stovepipes"

The inherent nature of XML schema enables the use of a consistent set of tags, yet does not prohibit adoption of multiple sets. Project managers implementing XML schema with metadata tags would be wise to collaborate so that interchangeable, metadata vocabularies facilitate what's being termed "semantic interoperability."

RECOMMENDATIONS

The potential described above, in small and large scale, requires the military to adopt an ontology, such as DAML. Systems must be equipped with applications that encapsulate system products (messages, data, overlays) with proper tags, and include metadata about each product. The systems must also be able to manipulate and process metadata from other systems. Once the semantic interoperability begins, new procedures for automatically sending and receiving filtered information should be installed with applications that process the information for specific purposes.

Project managers of legacy systems will need to adopt Army-wide XML schema, determine how to integrate the products of their respective system into an XML format with metadata for net-centric compatibility, and coordinate access of those products and/or applications to other systems.

The semantic web, with all of its promise for the military, requires systems to read, manipulate, process, and generate metadata. New applications, or agents, must translate derived data into an XML format and implement push/pull processes between echelons. New systems will need to integrate agents with existing system functionalities to allow remote procedure calls or method invocations from other connected systems. These capabilities are being developed in the business sector.

While the value of metadata is obvious as a means for improving search and retrieval processes, reducing bandwidth, and accessing information from a greater variety of files than ever, its greatest value lies in organizing and evaluating information.

CONCLUSIONS

The acquisition of decision superiority requires retrieving the right information at the right time. The existing Internet, with its reliance on HTML and limited <META> tags will not suffice. The Army's implementation of AKO represents a domain for transformation to a semantic web. If AKO maintains content while defining it within a hierarchical context, it will enable decision superiority. Metadata, explicit in label and implicit in structure, provides an interoperable channel for different systems to attenuate. Furthermore, applications that exploit the relationships between elements in the hierarchy can be used to manipulate, compare, and evaluate data. These "intelligent agents" will dynamically harness the store of knowledge and use it to bring decision superiority within our reach.

APPENDIX A. ACRONYMS AND ABBREVIATIONS

AKO	Army Knowledge Online
CoABS	Control of Agent-Based Systems
DARPA	Defense Advanced Research Project Agency
DAML	DARPA Agent Markup Language
DTD	Document Type Definition
HTML	Hypertext Mark-Up Language
IMS	Information Management System
LAN	local area network
MPEG	Motion Picture Experts Group
NIMA	National Imagery and Mapping Agency
NIST	National Institute of Standards and Technology
OPORD	Operational Orders
RDF	Resource Description Frameworks
SGML	Standard Generalized Markup Language
TASK	Taskable Agent Software Kit
TCP/IP	Transmission Control Protocol/Internet Protocol
TRADOC	Training and Doctrine Command
URI	Universal Resource Identifier
XML	Extensible Markup Language

APPENDIX B. TERMS

DAML: DARPA Agent Markup Language ties the information on a web page to machine readable semantics, created to provide tools and techniques to translate information from one machine to another, markup of reports to include semantics, and the use of agents.

HTML: Hypertext mark-up language is the computer language used to send text across the Internet and display information in a browser. Using predefined "tags," the display of text is defined by an author. The computer has no way to access the content of the text, just its display. Designed to link people to text, it is an application of the standard generalized mark-up language.

SEMANTIC WEB: technologies for enabling machines to make more sense of the web, with the result of making the Web more useful for humans.

SGML: Standard generalized markup language is a computer language designed for flexible marking of documents, through the use of tags. SGML is the base language that XML and HTML were written.

XML: application of SGML that allows custom definition of "tags." When data is tagged, it can be transferred across an Internet, WAN, or LAN. With proper implementation, (i.e. method of interpretation) interoperability between existing systems can be greatly extended.

APPENDIX C. REFERENCES

Berners-Lee, Tim, James Hendler, and Ora Lassila. "The Semantic Web." Scientific American. April 2002. 3 March 2003 <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&pageNumber=1&catID=2>

Berners-Lee, Tim. “XML And The Web.” Lecture 6 September 2000, XML World 2000, Boston, MA. Presentation online. 3 March 2003. <http://www.w3.org/2000/Talks/0906-xmlweb-tbl/>

Bray, Tim. "What Is RDF?” 24 Jan. 2001. O’Reilley XML. 3 March 2003 <http://www.xml.com/pub/a/2001/01/24/rdf.html>

Dornfest, Rael, and Brickley, Dan, "The Power Of Metadata.” 18 Jan. 2001. O'Reilly P2P, 3 March 2003. http://www.openp2p.com/pub/a/p2p/2001/01/18/metadata.html

Dumbill, Edd. " The Semantic Web: A Primer.” 1 Nov. 2000. O’Reilley XML. 3 Mar. 2003 <http://www.xml.com/pub/a/2000/11/01/semanticweb/index.html?page=2>.

Hunter, Jane. "Enhancing the Semantic Interoperability of Multimedia through a Core Ontology", accepted for IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description, Feb 2003. 3 March 2003 <http://archive.dstc.edu.au/RDU/staff/jane-hunter/events/paper.html>.

United States. Department of Defense. *IXO Overview.* Presentation. 1 August 2002. Defense Advanced Research Projects Agency (DARPA), Information Exploitation Office (IXO), Dr. Richard P. Wishner, Director. DARPATech 2002 Symposium. 3 March 2003. http://www.darpa.mil/DARPATech2002/presentations/ixo_pdf/slides/WishnerIXO_v4.pdf.

United States. Department of Defense. *Control of Agent Based Systems.* Presentation June 1999 at DARPATech ’99. Defense Advanced Research Projects Agency (DARPA), Information Systems Office, James Hendler, Systems Architecture and Integration Officer.

United States. Department of Defense. *Joint Vision 2020.* June 2000. Chairman of the Joint Chiefs of Staff, Office of Primary Responsibility: Director for Strategic Plans and Policy, J5; Approval Authority: General Henry H. Shelton, Strategy Division. U.S. Government Printing Office, Washington DC. 3 March 2003. http://www.dtic.mil/jv2020/jv2020.doc.

United States. Department of Defense. 8 August 2001. *Army Knowledge Management: A Strategic Plan for an Agile Force,* Version 2.1. Accessed 3 March 2003. <http://www.fgipc.org/02_Federal_CIO_Council/Downloads/61_ArmyKM_Strat_Plan.pdf >